Performance Analysis of Logistic Regression and Naive Bayes Classifiers Utilizing Different Datasets

Mert Gurkan, Mohammed Al-mekhlafi, Zainab Almheiri

***Abstract*—Logistic regression and naive Bayes are well-known machine learning algorithms that are widely used for various purposes. In this paper, we evaluate the performance and complexity of these two models on four different datasets. The datasets that we use to implement these models are Ionosphere, Adult, Wine quality, and Iris. Data preprocessing is a paramount step to obtain good performance of the implemented classifiers. We study the summary statistic of the observations, and underlying distributions of features in order to understand the nature and characteristics of the datasets. Besides, the objectives of this paper are, to implement logistic regression using (full batch) gradient descent to reach the optimal solution, and to implement naive Bayes using the appropriate likelihood for features (Gaussian and Bernoulli). Finally, we aim to compare the performance of naive Bayes and logistic regression on the four datasets based on accuracy and prediction time (speed). The results prove that both models can achieve good accuracy based on the characteristic of the given datasets.**

***Keywords*—Logistic Regression, Naive Bayes, cross validation, accuracy, complexity, features selection.**

I. Introduction

Machine learning is gained a lot of attention in the last decay in both academic and industrial. Machine learning is basically classified into supervised, unsupervised learning and semi-supervised. Supervised learning deals with labeled training data to perform learning that can be divided into regression and classification. The regressions aims to predict the quantity of the output. Classification method, on the other hand, is useful when an object needs to be assigned into a predefined group or class based on several observed attributes, or features, related to that object. Classifiers can be built with experience, which states data acquired from actual cases. The dataset can be preprocessing and expressed in a set of rules, such as it is often the case in knowledge-based expert systems, or serve as training data for statistical and machine learning models [1], [2].

Two of the widely used classifications machine learning technique are the Logistic regression and Naive Bayes. The learning and parameter estimation mechanisms are different between the two models. Naive Bayes is known as a generative classifier, whereas, the Logistic regression is a discriminative classifier. Generative classifiers (e.g., naive Bayes) learn a model of the joint probability, p(x, y), of the inputs x and the label y, and make their predictions by using Bayes rules to calculate $(p(y|x))$ and then picking the most likely label y. However, discriminative classifiers such (e.g., Logistic regression) model the posterior $(p(y|x))$ directly or learn a direct map from inputs x to the class labels [2], [3].

In this paper, we are interested in implementing Logistic regression and Naive Bayes and evaluating their performance on four datasets named, Ionosphere, Adult, Wine quality, and Iris (see section IV for detail data descriptions). We implement and analyse the accuracy of both Logistic regression and Naive Bayes classifiers. The results show that Logistic regression performs superior compared to Naive Bayes on a large dataset. Whereas, Naive Bayes performs well on a small dataset and its convergence time reaches its asymptotic within a shorter time.

The rest of the paper is organized as follows: Section II summarizes previous related work to the datasets that we use in this paper. Section III presents the analytical formulation of logistic regression and naive Bayes. In Section IV, datasets descriptions and preprocessing are provided. Section V presents the results. We conclude the paper in section VI.

II. Related Work

Previous studies have implemented different machine learning algorithms (classification based) on the applied datasets in this paper [4], [5]. In [4], a hybrid of naive Bayes and decision tree classifier, called NBTree was developed since Naïve Bayes does not perform well on large datasets. The results proved that the NBTree achieved high accuracy on the adult dataset. A feedforward neural network was applied on Ionosphere dataset, the results showed that model achived %100 accuracy on the training set and %98 accuracy on the testing set [5]. Additionally, a new algorithm called the neighborhood census rule (NCR) that is similar to the concept of k-NN classification techniques was proposed and applied on Iris datasets to predict the classes of iris plants. The results showed a good performance of the proposed method. Additionally, data mining approach was applied to predict wine quality using a big dataset called wine quality dataset. The dataset consists of white and Vinho Verde samples collected from Portugal. Three different algorithms were applied to this dataset, and the results proved that the support vector machine achieved promising results compared to the other methods.

III. Analytical Models

A. Problem Formulation

Classification machine learning aims to learn a mapping from inputs x to output 'y' where y includes values from $[1, C]$, C being the number of classes. If C is equal to 2, this is called as binary classification where we assume that y equals to 0 or 1. If C is greater than 2 then it is called as multi-class classification. [1], we apply two classifiers, Naive Bayes and Logistic regression to solve a binary classification to decide the class of y includes {0, 1} given a dataset observations D = (x_i, y_i) , i = 1, 2, ..., n of data items, where x_i is a vector of features

with known class of y_i. Besides, we implement Gaussian and Bernoulli Naive Bayes based on the likelihood of the features. This section summarizes the mathematical representation of the implemented model/classifiers.

B. Logistic Regression

Logistic regression is a generalized linear classifier that learns the probability of a sample belonging to a specific class. Logistic regression tries to find the optimal decision boundary that best separates the classes. The idea behind logistic regression is directly modeling the log-odds with a linear function given by Eq. (1). In logistic regression, no strong assumptions are made regarding the distribution of the explanatory variables. The cost function of logistic regression can be calculated using cross-entropy in Eq.(2). The optimal solution of parameters/weights w^*, on the other hand, can be estimated using the gradient descent algorithm where the weights are iteratively updated to minimize the cross-entropy in Eq.(3) [2] .

$$a = \sum_{i=0}^{m} w_i x_i = \mathbf{w}^T \mathbf{x} \tag{1}$$

$$J(w) = \sum_{n=1}^{N} y^{(n)} \log\left(1 + e^{-w^T x}\right) + \left(1 - y^{(n)}\right) \log\left(1 + e^{w^T x}\right) \tag{2}$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \sigma\left(\mathbf{w}^T \mathbf{x}_i\right)\right) \tag{3}$$

where α_k denotes the learning rate.

C. Naive Bayes

Naive Bayes is a classification method based on Bayes' theorem that derives the probability of the given feature vector being associated with a label under naive assumptions of conditional independence between features. It is appropriate for high dimensional data (high feature space). Additionally, the Naive Bayes model assumes that given a class G= j, the features (X_k) are independent, (see Eq.(4)). [6] In this paper, we apply Bernoulli and Gaussian Naive Bayes based on the underlying distribution of features, see Eq.(5) & (6), respectively.

$$f_j(X) = \prod_{k=1}^{p} f_{jk}\left(X_k\right) \tag{4}$$

$$p_{w_{\|d}}\left(x_d|y\right) = \text{Bernoulli}\left(x_d; w_{[d],y}\right) \tag{5}$$

$$p_{w_{[d]}}\left(x_d|y\right) = \mathcal{N}\left(x_d; \mu_{d,y}, \sigma^2_{d,y}\right) = \frac{1}{\sqrt{2\pi\sigma^2_{d,y}}} e^{-\frac{\left(x_d - \mu_{d,y}\right)^2}{2\sigma^2_{d,y}}} \tag{6}$$

In this equation, w_d takes one mean and standard deviation parameter for each class-feature pair.

	Ionosphere	Adult	Wine Quality	Iris
Missing values	No	Yes	No	No
Malformed features	2	3	N-	-
Duplicated instance	Yes	Yes	No	
Remove outliers	Yes	Yes	Yes	Yes
Categorization	No	Features	Output	Output
Normalization	No	No	Yes	Optional

TABLE I: Some of the Reprocessing steps

IV. Data Description and Preprocessing

A. Data Description

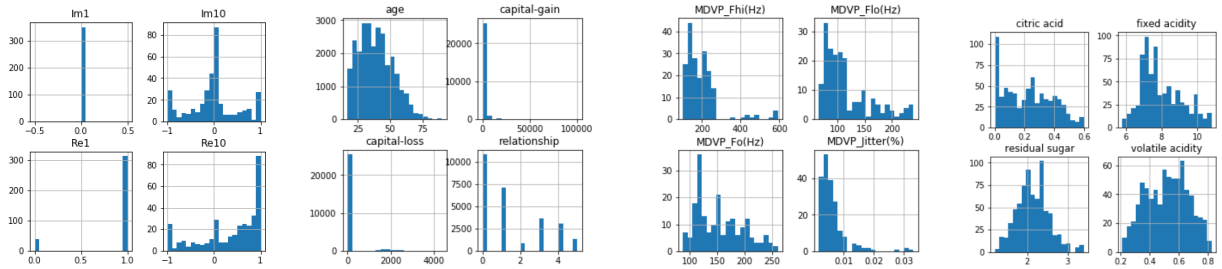
This section describes each dataset that we use to implement Naive Bayes and Logistic Regression classifiers the details of preprocessing is provided in next subsection:

1) Ionosphere dataset: consists of 351 instances with continuous 34 attributes represents 14 radar returns from the ionosphere to forecast the status of the weather is good or bad. The, correlation measurements are completed for the features. We get high correlation which 0.5, .3 and .2 for two a and three and eleven features respectively.
2) Adult dataset: is used to predict whether income exceeds 50K dollars per year or not, based given integer and categorical features. Adult dataset consists 48842 instances and 14 features. It includes features such as:*age, work-class, education, marriage-status, occupation, and sex.*. First, we clean the dataset for removing missing and duplicate values. We use one-hot encoding and categorization for the features.
3) Iris dataset: contains 3 classes of 50 instances each, where each class points to a type of iris plant. We use binary classification for this dataset. In order to implement this, we group classes Iris Versicolour and Iris Virginica. We use these two as one of the features and Iris Setosa as the second one.
4) Wine Quality dataset: The analysis determined of 13 continuous quantities. The size of dataset is 178 instances. Applying the normalization on this data set helps to improve the learning. To apply binary classification, A small modification is made on the quality deadset by making the quality of the wine is good if its quality more than 5 and bad otherwise.

B. Data preprocessing

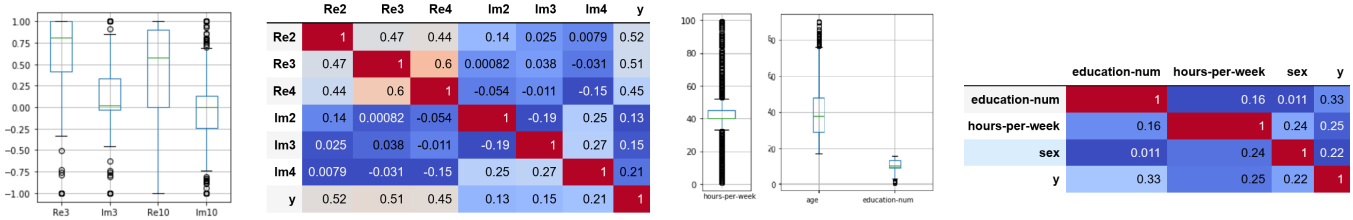
Data preprocessing is a paramount step to obtain good prediction results. This section discusses the procedures of data preprocessing in details as follows:

- We clean the datasets by removing missing values and spaces between observations from adult data set as the other datata sets are cleaned.
- We apply exploratory analysis to understand the statistical distribution of each feature, x_i, and remove malformed features, when more than 80 percent of the data points of a feature represents the same class. Fig. 1 shows the histogram of selected features for all the data sets. Fig. 3 presents the distribution of the aforementioned features to the output class.



(a) Ionosphere (b) Adult (c) Wine quality (d) Iris

Fig. 1: Histogram of selected features for the used datasets



	Re2	Re3	Re4	Im2	Im3	Im4	y
Re2	1	0.47	0.44	0.14	0.025	0.0079	0.52
Re3	0.47	1	0.6	0.00082	0.038	-0.031	0.51
Re4	0.44	0.6	1	-0.054	-0.011	-0.15	0.45
Im2	0.14	0.00082	-0.054	1	-0.19	0.25	0.13
Im3	0.025	0.038	-0.011	-0.19	1	0.27	0.15
Im4	0.0079	-0.031	-0.15	0.25	0.27	1	0.21
y	0.52	0.51	0.45	0.13	0.15	0.21	1

	education-num	hours-per-week	sex	y
education-num	1	0.16	0.011	0.33
hours-per-week	0.16	1	0.24	0.25
sex	0.011	0.24	1	0.22
y	0.33	0.25	0.22	1

(a) Ionosphere Boxplot (b) Ionosphere correlation matrix (c) Adult Boxplot (d) Adult correlation matrix

Fig. 2: Box plot and correlation matrix for selected features for both Adult and Inoshpere datasets

- We encounter that the first two features measured by the first radar are malformed features in Ionosphere dataset, while *capital loss, capital loss* and *race* race can be considered as malformed features.
- We remove outliers that greater than the standard deviation by three times.
- In Adult dataset, we remove redundant features such as *education*, and *fnlwgt* as it does not add any values to the model. We also convert some features such as *age, working class* into categorical features.
- We calculate the correlation coefficients as it is essential to study the association between each feature with the model output y .In Ionosphere dataset, y is highly correlated with only 11 features instead of the rest 32 features. Whereas, in Adult dataset, it is highly correlated with the *education-num and hours-per-week* as shown in Fig. 2 for selected features. Additionally, in Iris dataset, all feature are highly correlated with the model output. the same is applied for the wine dataset.
- We apply label encoder to label the output/class to be either $\{0$ or $1\}$. we develop one hot encoder to convert categorical features into binary.
- Finally, We randomly split the datasets into training and testing sets and we normalize them between $[-1, 1]$ during the training and the test phases. We create k-fold function for tuning the model hyper-parameters in order to control the learning process and reach the optimal solution. TABLE I summarize the important preprocessing steps based on the observations.

V. Results

A. Environment Settings

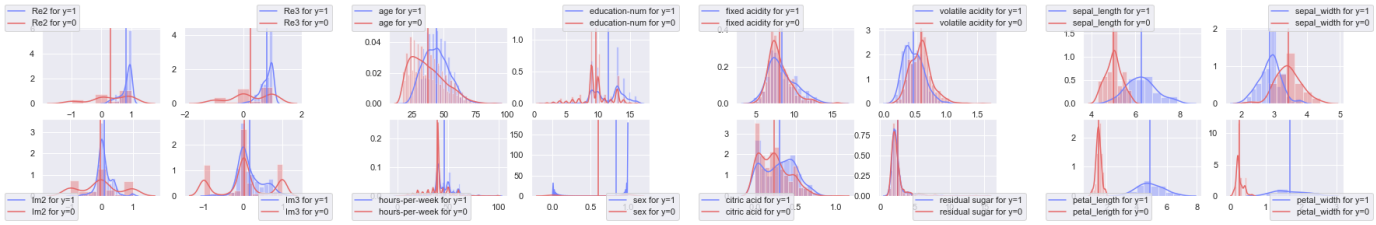
All the analysis was performed on Python 3.7 along with Pandas and Numpy libraries use the machine on with the following characteristics:

- System Type: x64-based PC
- Processor: Intel(R) i7-8700H CPU @3.20GHz,12 CPU
- GPU: NVIDIA Quadro P620.

B. Result of the model

Table II presents the performance in terms of the complexity and accuracy, of Logistic regression and Naive Bayes. Results show the superiority of Logistic Regression over Naive Bays in terms of accuracy and complexity. Both classifiers achieve the same accuracy, 100% accuracy on the testing set utilizing the Iris dataset as both classes are linearly separable. The prediction time for both classifiers is generally the same. However, Logistic regression takes a slightly longer time to coverage to its asymptotic. Excluding the adult dataset, Naive Bayes takes long for time convergence. This time is and not expected (even the complexity is increases due large number of features used and the binary nature of the features).

Fig. 4 depicts the model performance of Logistic regression on the utilized datasets including Ionosphere (a), Adult (b), Iris (c), and Wine (d) versus the number of iterations at different learning rate and penalty term (lambda). Results show that the model performance increases with the increase of the number of iterations,(the accuracy converges slower when the learning rate decreases) at the learning rate equals to 0.1, and lambda between 0 and 0.001. Additionally, it is worth mentioning that the curve of the Adult dataset (b) is smoother than others since it includes categorical (mapped to binary) features. The

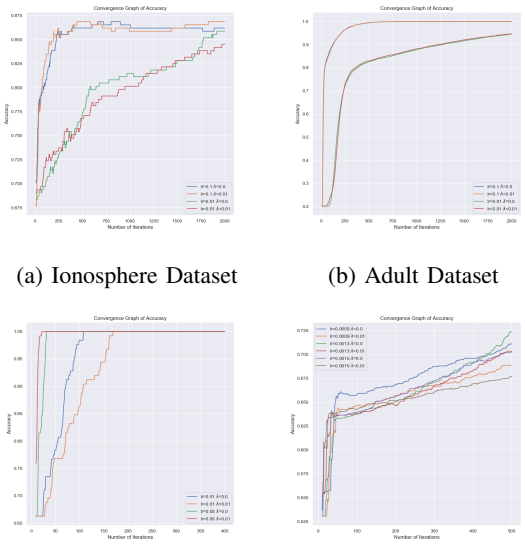


(a) Ionosphere (b) Adult (c) Wine quality red (d) Iris

Fig. 3: Box plot and correlation matrix for selected features for both Adult and Inoshpere datasets

	Ionosphere		Adult		Wine Quality		Iris	
	LR,$\alpha = .01$	GNB	LR,$\alpha = 0.1$	GNB	LR,$\alpha = .01$	GNB	LR,$\alpha = .0001$	GNB
K-fold Time (sec)	0.3241	0.0010	3.243	.008	0.8407	.01	0.2862	0.0079
K-fold Accuracy %	90.85	81.70	100	80.39	62.72	45.77	100	100
Test Time (sec)	0	0.005	0.001	1.062	0	.0159	0.01	0.001
Test Accuracy %	90.02	85.37	100	79.7	61.53	43.95	100	100

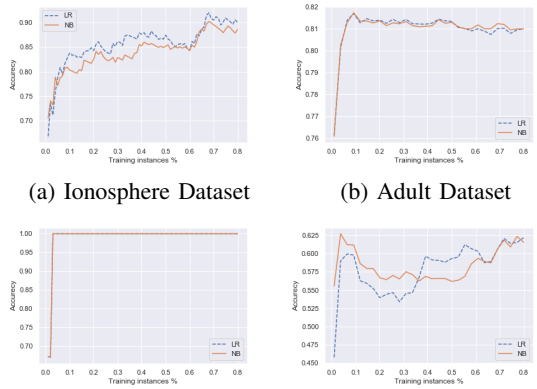
TABLE II: Performance and Complexity analysis of the Logistic Regression (LR) and Gaussiane Base (GNB) classifiers. For LR, the stopping cost is set to be 10^{-5}



(a) Ionosphere Dataset (b) Adult Dataset

(c) Iris Dataset (d) Wine Dataset

Fig. 4: Performance



(a) Ionosphere Dataset (b) Adult Dataset

(c) Iris Dataset (d) Wine Dataset

Fig. 5: Performance

Iris is linearly separable so it converge with small number of iterations.

Fig. 5 depicts the model performance of Logistic regression and Naive Bayes versus the number of training instances after cross-validation procedures and selecting hyper-parameters. Results show that Logistic regression performs better than Naive Bayes with the increase of training instances when utilizing the Ionosphere dataset. However, both classifiers achieve the same performance on the adult dataset. Similarly to the Iris dataset, and surprisingly both classifiers achieve the same accuracy (100%) regardless of the number of training instances. However, the performance of Logistic regression is better on small wine dataset compared to Naive Bayes while the prediction time for both classifiers is generally the same.

VI. Conclusion

In this paper, we evaluate the performance and complexity of the logistic regression and Naive Bayes on four datasets including Ionosphere, Adult, Iris, and Wine. Overall, Logistic regression shows superiority over Naive Bayes on large size datasets. However, both classifiers achieve the same performance on Adult and Iris datasets. We find that the performance of a classifier depends on the nature and characteristics of the dataset. Logistic regression accuracy increases as the number of iterations increases. we also find that the learning time of Logistic regression is longer compared to Naive Bayes. Additionally, we conclude that as the size of training instances increases the model performance increases. For future work, we recommend to apply a unique feature selection method such as principle component analysis (PCA).

VII. Statement of Contribution

Mert contributed to upload and select the datasets and write them in write-up. Mohammed and Zainab contributed to writing the code for the remaining parts and adding them to the write-up.

REFERENCES

[1] C. Robert, "Machine Learning, a Probabilistic Perspective," 2014.

[2] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[3] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in neural information processing systems*, 2002, pp. 841–848.

[4] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in *Kdd*, vol. 96, 1996, pp. 202–207.

[5] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," *Johns Hopkins APL Technical Digest*, vol. 10, no. 3, pp. 262–266, 1989.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.